REJEPA: A Novel Joint-Embedding Predictive Architecture for Efficient Remote Sensing Image Retrieval

Supplementary Material

1. Introduction

We provide additional qualitative analyses to further validate the effectiveness of REJEPA for remote sensing image retrieval.

- In Section 2, we present an extensive comparison of REJEPA against existing methods on more challenging query images, demonstrating its robustness in retrieving semantically relevant results across diverse scenarios. These comparisons reinforce the model's ability to handle high intra-class variance and complex spatial patterns.
- In Section 3, we analyze the attention heatmaps of REJEPA's context encoder, showcasing its superior ability to focus on structurally significant regions compared to pixel-reconstruction-based models. This visualization confirms that feature-space prediction not only enhances semantic representation but also achieves better retrieval performance with significantly lower computational overhead.

2. Qualitative Performance

To further validate the retrieval effectiveness of REJEPA, we perform qualitative comparisons against SatMAE and SatMAE++ on challenging query images, focusing on crop fields, a crucial application in remote sensing content-based image retrieval (RS-CBIR). Accurate retrieval of semantically similar agricultural regions is essential for monitoring crop health, assessing land use changes, and analyzing agricultural patterns over time.

Figure 1 presents retrieval results for crop field queries across the FMoW datasets. REJEPA consistently retrieves images that exhibit structural and spectral similarities to the query, preserving fine-grained texture and spatial distribution. In contrast, SatMAE[1] and SatMAE++ [2] often retrieve images with visually dissimilar characteristics due to their reliance on pixel-level reconstruction, which struggles to capture high-level semantic relationships. Notably, REJEPA achieves greater consistency in identifying homogeneous vegetation areas, correctly retrieving fields with similar crop structures and spectral signatures.

Advantages of Feature-Space Prediction: Unlike pixel-reconstruction-based models, which emphasize low-level pixel alignment, REJEPA learns feature-space representations that prioritize semantic information. This enables robust retrieval across varying illumination conditions, seasonal changes, and different geographical locations, making it particularly suitable for large-scale agricultural monitor.

3. Attention Visualisation

To further demonstrate the effectiveness of REJEPA, we present attention heatmaps from the context encoder and compare them with those of SatMAE and SatMAE++. These visualizations provide insights into how feature-space prediction enhances representation learning, leading to more discriminative and semantically rich retrieval performance.

Comparison of Attention Patterns: Figure 2 and 3 shows attention maps for representative remote sensing scenes, including urban areas, agricultural fields, and infrastructure sites. The attention heatmaps reveal distinct differences between REJEPA and pixel-reconstruction-based models:

- **REJEPA:** Focuses on high-level semantic structures such as roads, buildings, and vegetation clusters, effectively capturing scene-relevant spatial patterns.
- **SatMAE and SatMAE++:** Display scattered and less interpretable attention distributions, often failing to capture coherent object structures due to their reliance on pixel reconstruction.

References

[1] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 1

[2] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27811–27819, 2024. 1

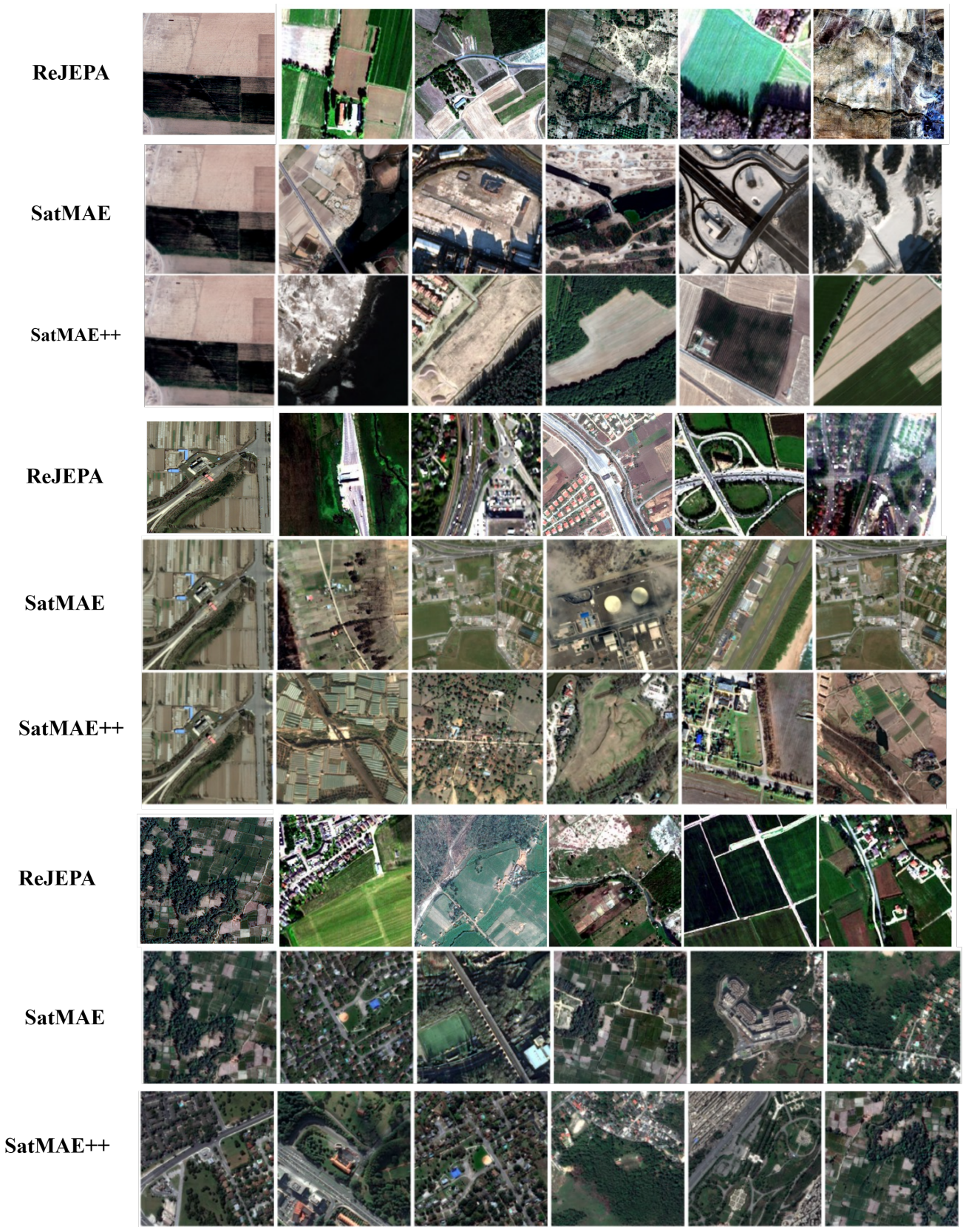

Figure 1. Comparative performance of REJEPA on critical queries with SatMAE and SatMAE++

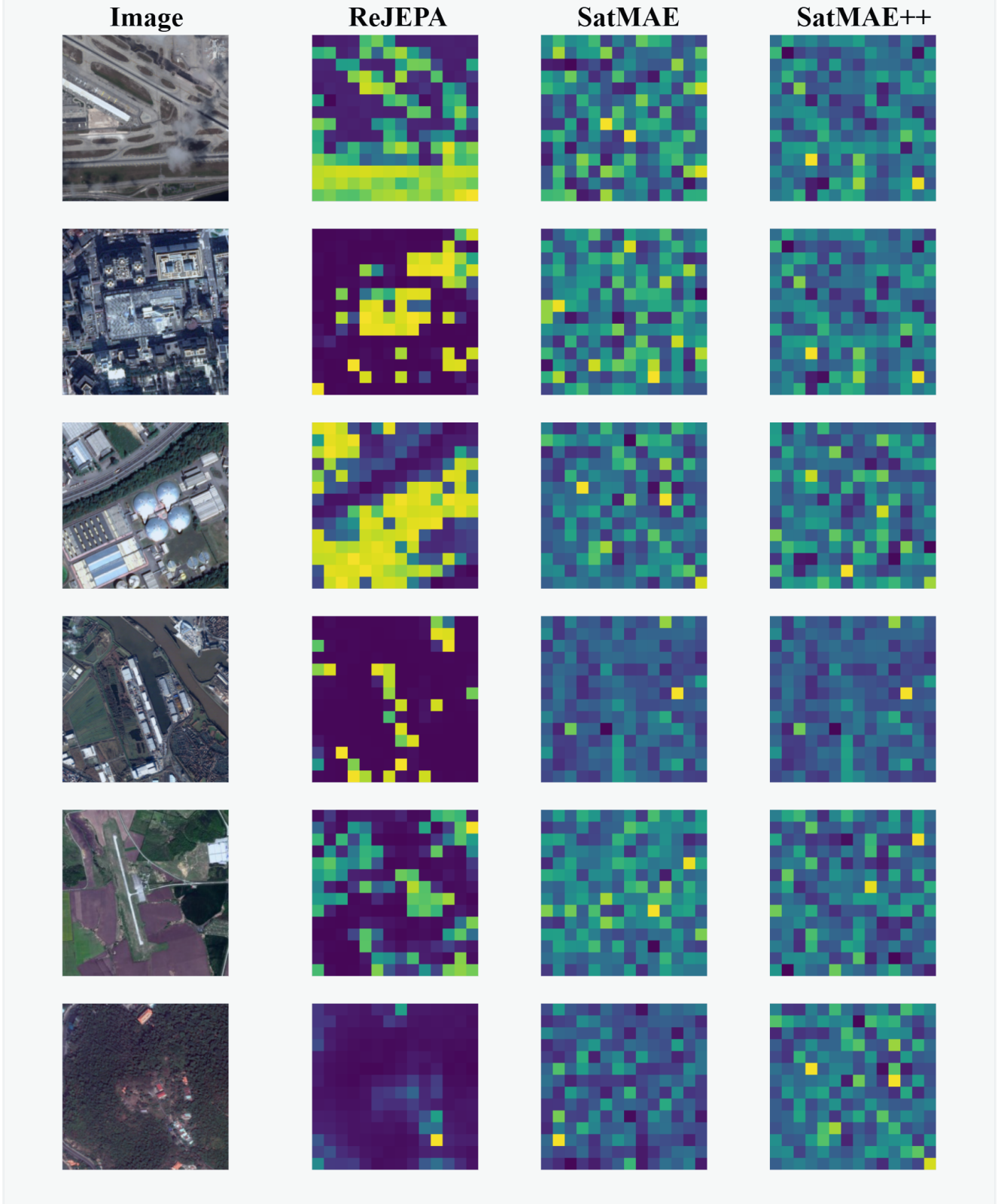

Figure 2. Attention heatmaps of the context encoder REJEPA vs the pixel-wise reconstruction encoders of SatMAE and SatMAE++

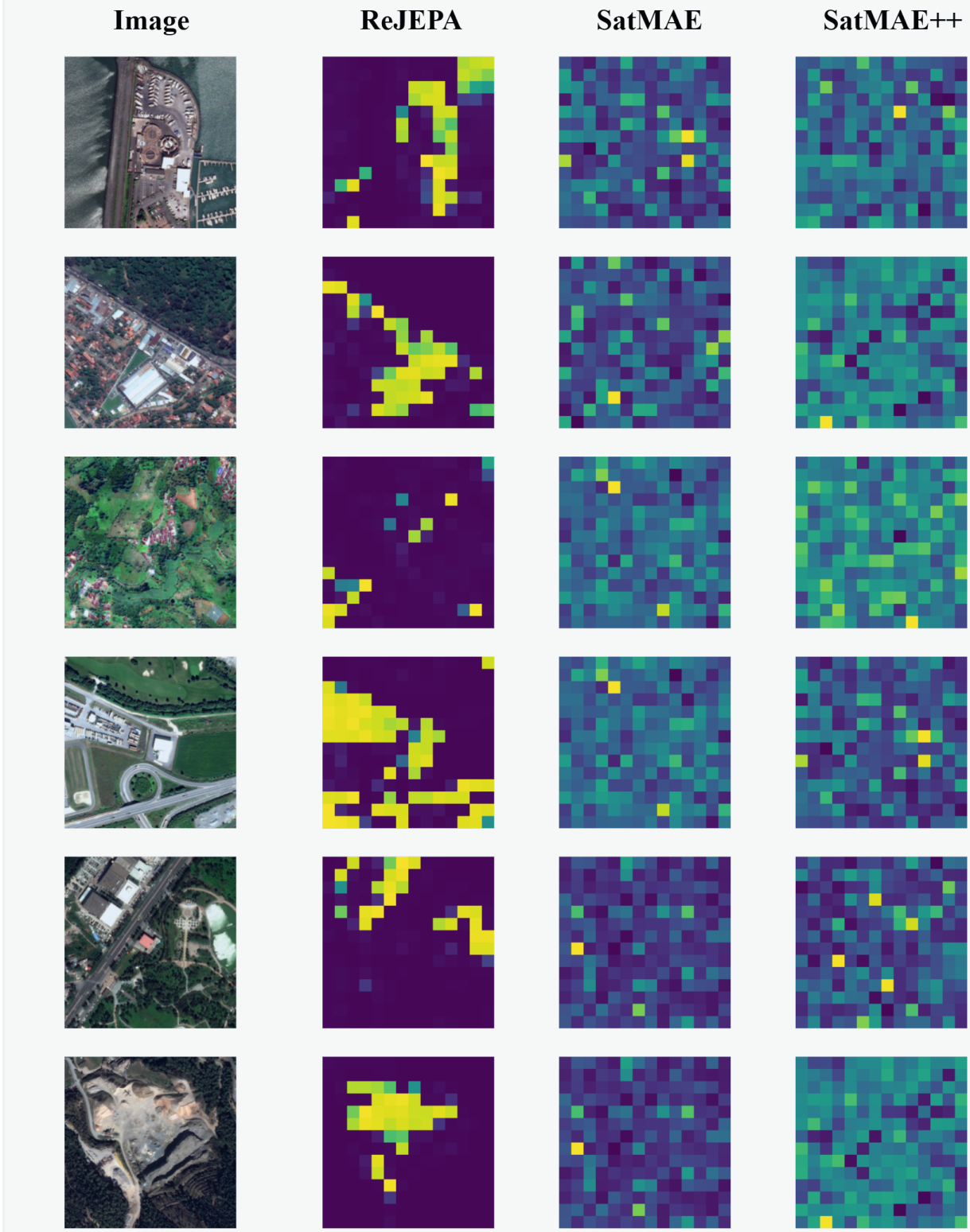

Figure 3. Attention heatmaps of the context encoder REJEPA vs the pixel-wise reconstruction encoders of SatMAE and SatMAE++